Webページとしての類似性を利用した Linked Data リポジトリの自動収集手法
A method for collecting Linked Data repositories based on similarities of webpages associated with them

学籍番号：201321637
氏名：瀬尾　崇一郎
Soichiro SEO

近年、Web を通じたオープンデータの動きが盛んになってきている。このオープンデータについて、Linked Data と呼ばれる実践的方法によって行う Linked Open Data と呼ばれる公開方式が W3C によって推奨されており、様々な機関や団体が自らの Linked Data を提供するための Linked Data リポジトリを公開している。利用者が Linked Data リポジトリを発見するための方法として現在主として使われているのはデータカタログサイトと呼ばれるポータルサイトである。これにはリポジトリ情報の登録や更新を人手の作業に頼っているという問題点があり、リポジトリ数が世界中で増加し続ければ対応できなくなってしまうことが予想される。そこで本研究では、Web 上で公開されている Linked Data リポジトリを自動的に収集するための手法を提案する。

本研究では Linked Data リポジトリを収集する方法として、検索エンジンが Web ページ収集のために稼働させているクローラを利用する。多くの Linked Data リポジトリは、SPARQL Endpoint と呼ばれる Web API をブラウザから利用するための Web ページ（以降、Web UI）を用意しており、検索エンジンのクローラはこれも収集していると考えられる。また Web UI は利用ソフトウェアなどといった構築環境の理由から他のリポジトリの Web UI と Web ページとして類似することが多い。本研究ではこの性質を利用し、既知の Web UI を類似度に基づいてクラスタリングすることで類似 Web UI 群を判別し、その類似 Web UI 群ごとに機械的に抽出した特徴的フレーズを用いることで、クローラ型検索エンジンを利用した Linked Data リポジトリの自動収集を可能とした。

この提案手法について、実際にサンプルとして the Datahub から取得した Linked Data リポジトリを使用し、自動収集の実験を行った。その結果、提案手法によって類似 Web UI 群ごとの特徴的なフレーズを抽出することができ、この特徴的フレーズを Google 検索のフレーズ検索に利用することで、the Datahub に登録のある既知のリポジトリおよび登録のない未知のリポジトリを収集できることを確認した。更に特徴的フレーズは類似 Web UI 群ごとに複数を抽出し併用することで、収集能力がより向上することを発見した。

研究指導教員：阪口　哲男
副研究指導教員：永森　光晴